THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION

MEASUREMENT ERRORS IN SIPP PROGRAM REPORTS

No. 113

K. H. Marquis and J. C. Moore
Bureau of the Census

U. S. Department of Commerce BUREAU OF THE CENSUS

TABLE OF CONTENTS

MEASUREMENT ERRORS IN SIPP PROGRAM REPORTS

Kent H. Marquis and Jeffrey C. Moore
U.S. Census Bureau

ABSTRACT

An administrative record check of program participation reporting in the first two SIPP interviews shows that, while response errors are rare, they have important biasing effects on estimates of means and correlations. Our search for models of the causes of the errors includes classical hypotheses about forgetting, memory decay, confusion about name attributes, telescoping, learning to underreport, interviewers, and proxy responses. While these hypotheses give occasional insights into isolated error problems, they do not provide a fundamental understanding of the error dynamics. We mention some exploratory cognitive research that may provide the broader understanding and may be useful in devising better measurement procedures.

KEYWORDS

Response errors, Memory decay, Telescoping, Interviewer variance, Proxy reporting, Cognitive heuristics

1. INTRODUCTION

Information from government surveys is important both to policy planning and to a basic understanding of how society functions. But if we want to inform our policies and theories with survey data, then the data should either be accurate or we should have a thorough knowledge of their error structure. A record check is a good way to get descriptive information about accuracy and errors. And from good descriptions may come good solutions to the problems that cause the errors. This paper describes the results of a record check study for the Census Bureau's Survey of Income and Program Participation (SIPP).

In Section 2 we discuss SIPP, our record check study methods, and our response error estimates. Then, in Section 3, we look at the descriptive results, seeing that while response errors are very rare, they have important distorting effects on estimates that analysts make from SIPP data. In Section 4 we describe our tests of some classical hypotheses about the causes of the response errors including models of forgetting, confusion, learning and competence. These models are not especially useful in explaining the response errors. We discuss some implications of the results in Section 5 and, on the basis of some new cognitive research, suggest the general direction that a future research program might take to describe and reduce response errors in SIPP.

2. BACKGROUND AND METHODS

2.1 SIPP

SIPP is a longitudinal panel survey designed to provide improved information about the economic situation of people and families in the United States. For each person fifteen years of age or older, SIPP collects monthly information about earnings, participation in government transfer programs, assets and liabilities, labor force participation, and related topics, for the four months preceding the interview month. Generally, a panel consists of eight such interviews, covering about 2 1/2 years. Proxy reporting is permitted for household members not available for interview at the time of the visit. For a detailed description of the SIPP program, see Nelson, McMillen, and Kasprzyk (1985).

2.2 The SIPP Record Check Study Design

The purposes of the SIPP Record Check Study are to provide an evaluation of the quality of the major program participation data gathered in SIPP and to generate ideas for improving the data quality. Elsewhere (Moore and Marquis, 1989) we have described the project in detail. Below we summarize the major aspects of the research, including the record check design; the people, programs, and time periods which comprise the data for the study; and the matching procedures employed.

2.2.1 Basic Record Check Design

The SIPP Record Check uses a "full" rather than a one-directional design, which permits the evaluation of the full range of survey responses--for example, both "yes" and "no" reports of program participation. Marquis (1978) describes the limitations of partial designs (e.g., checking records only for those who report in the survey that they possess the characteristic of interest; or surveying people known to possess the characteristic to see if they report it), which are almost guaranteed to produce biased estimates of survey measurement errors. Full designs are necessary for producing unbiased estimates of the parameters of the response error distribution.

2.2.2 Programs

We obtained program participation records for eight government transfer programs, half administered by the states and half administered by the Federal Government. These programs, and their acronyms are:

<u>State-administered programs:</u>	
Aid to Families with Dependent Children	(AFDC)
Food Stamps	(FOOD)
Unemployment Insurance	(UNEM)
Workers' Compensation	(WORK)
<u>Federally administered programs:</u>	
Federal Civil Service Retirement	(CSRET)
Old Age Survivors Disability Insurance ("social security")	(OASDI)
Supplemental Security Income	(SSI)
Veterans' Pensions and Compensation	(VETS)

From each agency we obtained identifying information (for matching) and monthly benefit receipt information (for response error assessment) for all persons who received income from the target program at any time from May 1983 through June 1984 (see below). The administrative records provide comprehensive coverage of the population in each state, and define program participation and benefits in virtually the same way that SIPP does.

2.2.3 Time periods

The interview data are from the first two interviews ("waves") of the 1984 SIPP Panel, for which interviewing began in October 1983. Figure 2.1 illustrates the wave, rotation group, interview month, and reference period structure for the survey data. As shown in the figure, the calendar months in the reference periods for the first two

Wave	Rotation Group	Reference Period Months: 1983 Jun	Jul	Aug	Sep	Oct	Nov	Dec	1984 Jan	Feb	Mar	Apr	May
1	1	4	3	2	1	(I)							
	2		4	3	2	1	(I)						
	3			4	3	2	1	(I)					
	4				4	3	2	1	(I)				
2	1					4	3	2	1	(I)			
	2						4	3	2	1	(I)		
	3							4	3	2	1	(I)	
	4*/								4	3	2	1	(I)

KEY: (I) = interview month
Reference Period: 4--3--2--1 = 4 months ago, 3 months ago,...,last month

*/ Technically, rotation group 4 was not administered a wave 2 interview. The "missing" interview was transparent to respondents who simply received their wave 3 interview at the time they would have received the wave 2 interview. All references in this paper to "wave 2" include the wave 3 interview for this portion of the panel.

Figure 2.1: Survey Structure for Data Included in the SIPP Record Check Study

interviews for all rotation groups include June 1983 through April 1984. In our analyses, however, we ignore calendar months, and instead refer to the time periods covered by the survey data in terms of SIPP wave and reference month--e.g., wave 1, month 4; wave 1, month 3, etc. This is preferable because of the staggered rotation group structure of SIPP.

2.2.4 States and People

We conducted the record check study in four states: Florida, New York, Pennsylvania, and Wisconsin. These states were selected for convenience, and are not necessarily representative of the larger SIPP sample. The primary selection criteria included the following:

1) a reasonably large SIPP sample;

2) an appropriate, high quality, computerized, comprehensive, and accessible administrative record system for the programs of interest;

3) a willingness to share detailed, individual-level data for purposes of the research; and

4) some geographic diversity.

For the first two waves of the 1984 SIPP Panel the total SIPP sample included about 20,000 interviewed households. Of these, about 5,000 were included in the record check. And about 11,000 people lived in the record checked households.

The analyses reported in this paper do not use all available SIPP sample persons. The major restriction is that the approximately 2,700 children under age 15--who are included as sample persons but not interviewed--are excluded. Other restrictions are as follows:

1) approximately 350 adult sample persons who refused to report their social security number in the survey (SSN refusers) were excluded from the personal identifiers file made available to us for matching--although we have survey data for these people, we exclude them from our analyses because they were not subjected to matching against the administrative records;1/

2) approximately 500 adult sample persons for whom data reported by self or proxy were not available for all eight months (e.g., deaths, movers, refusers) are excluded from the analysis files; and

3) for the state-administered programs (AFDC, FOOD, UNEM, and WORK) we exclude the New York portion of the sample, about 2,300 cases, because there are some unresolved issues concerning the quality of selected data fields in the available New York administrative files.

For the Federal-level programs, then, the total number of sample persons available for analysis is about 7,550; for the state-level programs about 5,200.

2.2.5 Matching

We used the computerized matching software developed by the Census Bureau's Record Linkage Research Staff (e.g., LaPlant, 1989, Jaro, 1989), which is based on the theoretical work of Fellegi and Sunter (1969). The major advantages of this system (over, say, a clerical match) are its speed, its ability to process huge data sets, its ability to evaluate a match based on many variables simultaneously, and its ability to resolve, consistently and objectively, possible matches that differed on the value of one or more match variables. We matched on variables that were very likely to uniquely

---

1/ Occasionally our matching procedures matched an SSN refuser's administrative record(s) to a child in the same household. When we deleted children from the current analysis group, we attempted to rematch to an adult in the same household any administrative record(s) previously linked to the child, using whatever match information was available. If we judged that the match was "good," we relinked the administrative record information to the new person. Otherwise we did not link the administrative record information to anyone retained in the analysis group. A "good" match was one where there was better agreement on available match information such as name, age, sex, etc. Thus, a small number of SSN refusers are reincluded into the analysis group for selected programs (usually not more than two or three per program).

identify people such as their name, address, social security number and date of birth. See Moore and Marquis (1989) for a description of the matching techniques used in the record check.

2.3 Definition of Response Errors

In this paper we estimate errors in reports of program participation, a binary variable where 0 means not participating and 1 denotes participation (in the sense of receiving benefits from the program). The response error scores are derived by comparing responses from SIPP to the true values from administrative records. We discuss several kinds of response error, all defined from the 2 x 2 table in Figure 2.2.

REPORTED PARTICIPATION	TRUE PARTICIPATION YES = 1	NO = 0	
YES = 1	a	b	
NO = 0	c	d	
	a + c	b + d	N

Figure 2.2: Notation for Cross-Classified Reported and True Values.

The letters a, b, c, and d in the table represent frequencies of reported and true characteristics. N is the sample size.

The total number of WRONG ANSWERS (or misclassification errors) for a program is b + c. The misclassification rate is
(b + c) / N
and the misclassification percent (or percent wrong) is
[(b + c) / N] x 100.

The frequency of UNDERREPORT errors is c. The underreporting error rate, which is conditional on a true positive, is c / (a + c), and the percent of underreporting errors is 100 times the rate.

Similarly, the frequency of OVERREPORT errors is b, the rate is b / (b + d), and the percent is 100 times the rate.

We will use the percent wrong in Section 3 (descriptive results) and reserve the underreport and overreport statistics for Section 4 (hypothesis testing results).

2.4 Descriptive and Inferential Statistics

For each program, we usually calculate descriptive statistics (e.g., percent wrong) for each month and report an average over the entire eight months (or other groups of time periods such as wave 1 and wave 2). Unless we say otherwise, the inferential statistics refer to these averages. For the hypothesis tests and other "within person" comparisons, most inferences are based on paired-comparison t-tests that take into account the correlation of the observations for each person over time. We reject the null hypothesis for $p \leq .05$. We discuss other inferential procedures as they are used. For all of our inferential statistics we assume simple random sampling although the SIPP sample design is more complex than this. As a result, our population variance estimates and corresponding p-values are likely to be slightly underestimated for the individual monthly or program-specific analyses. However, we feel that our stated conclusions, based on consistent patterns across programs and time periods, would not change if we were to take the complex sample design into account in our variance estimates.

We call the effect of response errors on a parameter estimate a bias. The bias is the difference between the parameter estimated with data containing response errors and the true parameter value. We will examine two kinds of parameter estimates, a mean and a correlation. The bias in the estimated mean is {[(a + b) / N] - [(a + c) / N]} or (b - c) / N. Dividing by (a + c) / N yields the percent bias. In the Appendix we derive the percent bias for the correlation estimate, the statistic we use in this paper. In an earlier paper (Marquis and Moore, 1989b) we also derived the expressions for the bias in two forms of the bivariate regression coefficient estimate. However, the correlation result is a reasonably good summary of the two regression results.

3. DESCRIPTIVE RESULTS

In this section we will look at the response error percents for measures of program participation level and change. And we will examine the effects of those errors on statistics that analysts estimate using SIPP data. While the percentage of responses in error is always very small, the errors have moderate to large effects on estimates.

3.1 Misclassification Rates

The misclassification error percentages for monthly reports of program participation or, more simply, the percentages of wrong answers, are very low for each of the eight SIPP programs in the record check study. In Figure 3.1 we average over the eight months of data to look at the percent wrong in reporting participation level. We observe that the lowest error rate is 0.2 percent (for CSRET) while the highest is 2.3 percent (for OASDI). Thus, response errors are extremely rare regardless of which program is involved.2/

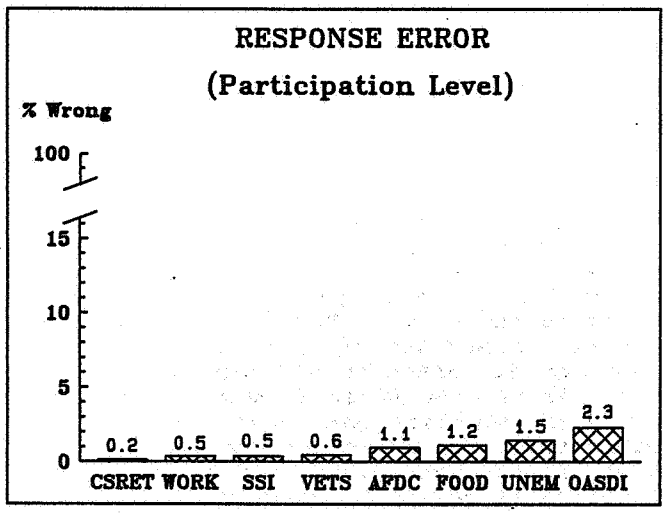

Figure 3.1: Average response error percentages for program participation are very low.

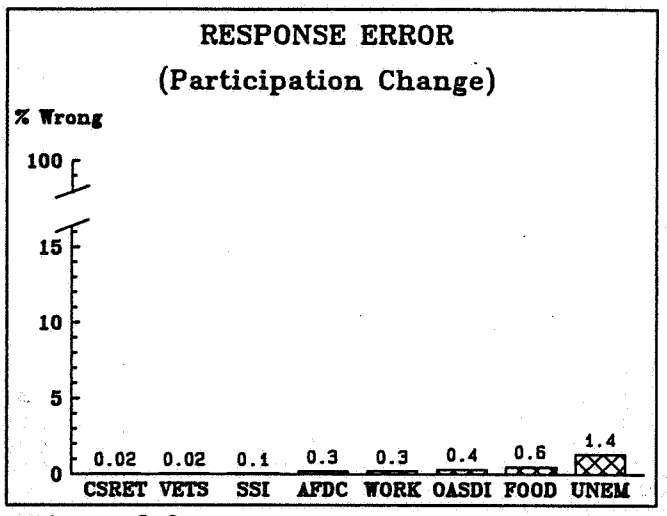

Figure 3.2: Average response error percentages for participation change are also very low.

Next, let us look at the percent response error in measures of program participation change. For any two adjacent months, we say a change has occurred when the program participation status is different (yes in one month and no in the other month, ignoring the direction of change). If the participation status is the same (either both yes or both no), then we say that no change has occurred. In Figure 3.2 we have averaged the percent wrong in change measures over the seven possible pairs of adjacent months and we see even lower error rates. They range from .02 (two-hundredths) percent for CSRET to 1.4 percent for UNEM. So errors in measures of starting or stopping the receipt of program benefits are also very rare.3/

Put another way, almost all respondents report participation in each of the tested programs accurately almost all of the time.

3.2 Effects of Response Errors on Estimates

Now we ask whether these low response error percents make much difference in the statistical estimates that subject matter analysts might make from SIPP data. If the effects are small, then response error reduction should not be a major concern in the SIPP program. On the other hand, if the effects are large, then it is important to bring the errors under control as quickly and completely as possible.

We will look at the biases induced by response errors in two kinds of estimates: the mean and the correlation. The mean estimate could be something like the proportion of the sample enrolled in the Food Stamps program in the month of June. An example correlation estimate might be between education level and participation in the Food Stamp program last month. In deriving the correlation bias, we assume that the program participation variable is measured with error and that the other variable is perfectly measured. This allows us to show the "pure" biasing effect of measurement error in the

---

2/ For interested readers, we show in Appendix Table 1 both survey-reported and true monthly program participation for all programs and months.

3/ Again, interested readers will find in Appendix Table 2 both survey-reported and true month-to-month program participation change data for all programs and time periods.

participation variable. For both the means and the correlations, we made separate estimates of bias for each of the eight (or seven) time periods and report the average of the monthly biases here.

3.2.1 Effects of Errors on Mean Estimates

Figure 3.3 shows the bias in estimates of the level of program participation. The net bias is usually negative for this sample, indicating that the estimated mean is usually lower than the true mean when using the SIPP data containing response errors. Biases for some programs, such as VETS and OASDI are trivial, only minus three percent and plus one percent. But for other programs, such as the 18 percent underestimate for WORK and the 39 percent underestimate for AFDC, the biases are more serious.

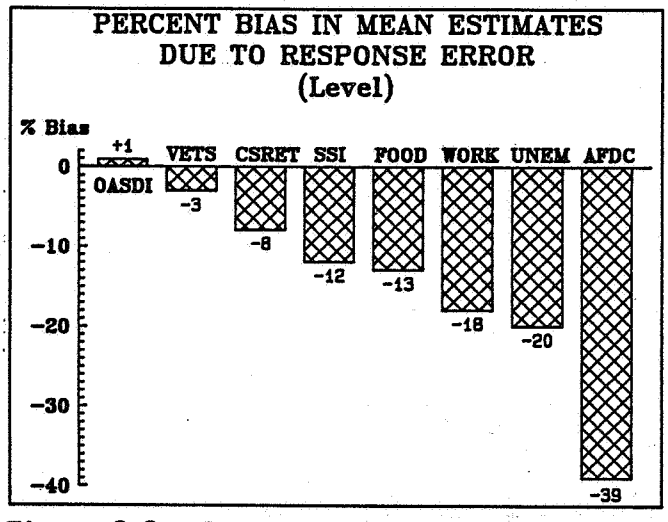

Figure 3.3: Response errors usually bias estimates of program participation levels in a negative direction.

Turning next to the biases in estimates of mean change rates for program participation, we first introduce the concept of the "seam" between interviews, since prior research suggests that the biases may be affected by this timing indicator. Recall that a change refers to whether program participation is the same or different in any two adjacent months. If the two adjacent months are reported in two different interviews, we refer to that time period as "on the seam" between the two interviews, and a change in this period is called an on-seam change. Change measured in any other pair of adjacent months is an off-seam change. This is illustrated in Figure 3.4.

Previous research (Moore and Kasprzyk, 1984; Burkhead and Coder, 1985; Hill, 1987) indicates that many more changes are measured on the seam compared to off the seam. This is also true for this sample as we show in Figure 3.5. Take, for example, the middle data for the Food Stamps program (FOOD): even though we would expect the rates

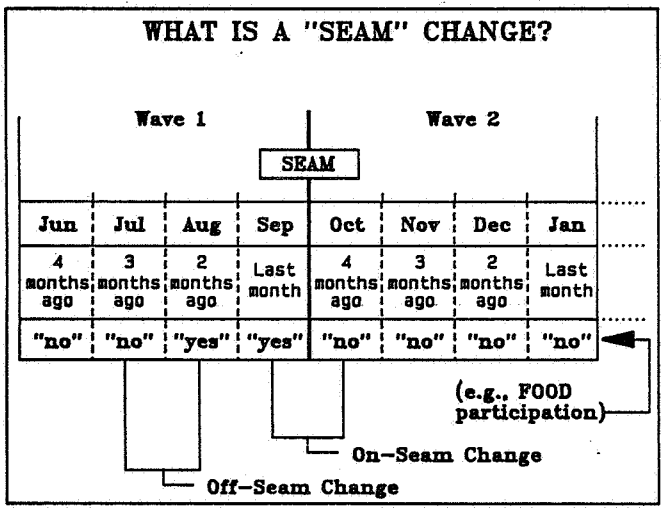

Figure 3.4: A program participation change is "on seam" when it occurs across months covered by different interviews; "off seam" changes occur across months within the same interview.

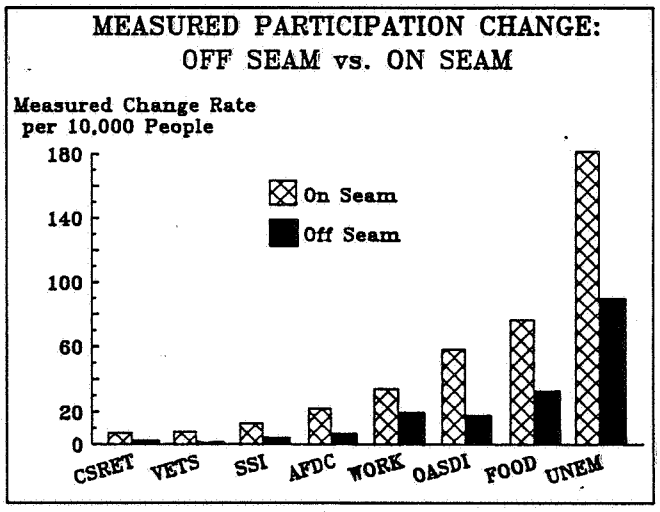

Figure 3.5: Much more change is measured "on seam" than "off seam."

to be the same, respondents reported change at the rate of 77 per 10,000 people on the seam and at the much lower rate of 32 per 10,000 in the average pair of off-seam months. This pattern is repeated for each of the other programs also. All of the on-off seam differences are statistically significant. So we turn, now, to the record check data to determine which of these estimates is correct, the on or off seam estimate. The results are surprising since neither estimate is generally correct.

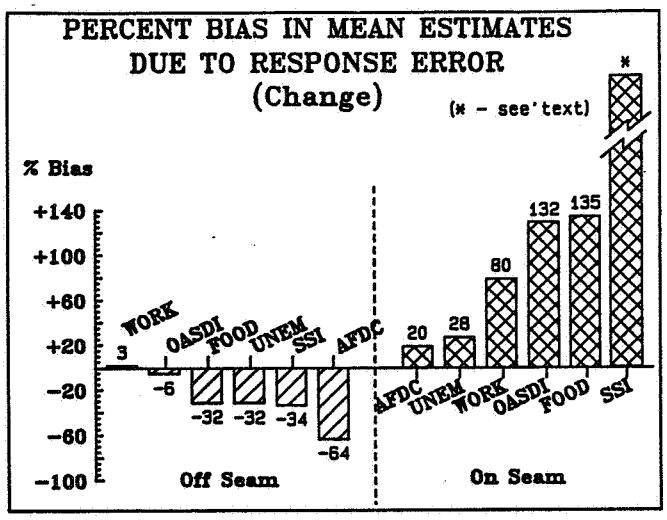

Figure 3.6: The sign of the change bias due to response error depends on whether change is measured on or off the seam.

Looking, in Figure 3.6, at the effects on mean change rate on and off the seam, we see that almost all of the off-seam biases are negative and all of the on-seam biases are positive. Thus, too few program participation changes are measured for the off-seam months and too many inferred for the on-seam months. The size of the on-seam bias estimate for SSI is especially uncertain due to a true change rate that, by chance, was abnormally low for the seam time period. Imputing an expected true change rate, based on true change rates for the other month pairs, the new bias estimate would be about 200 percent instead of 900 percent as originally estimated. We have omitted estimates for the two of the eight programs because their true change rate in at least one pair of months was zero (a + c = 0), so the percent bias could not be determined.

(Some may wonder whether the total number of changes is over-, under- or accurately estimated over the two waves. The results--not shown--do not follow a simple pattern. Total change is overestimated in some instances, underestimated in other instances and, in still other cases, the estimated total comes close to the true total.)

Next, we will look at the effects of the response errors on correlations. These results show very different patterns.

3.2.2 Effects of Errors on Correlation Estimates.

As shown in Figure 3.7, the effect of response error is to attenuate the bivariate correlation estimate, causing it to move closer to zero than the true value in the sample. We derive our estimate of the bias in the correlation in Appendix 1. The correlation is between the reported participation status in a month and a variable that is assumed to be measured without error, in order to focus attention only on the one set of measurement errors. Results indicate small to moderate percentages of bias for the first five programs (20 percent or less) and moderate to large attenuation for the remaining programs (33 to 51 percent). These effects can cause even the skilled analyst substantial trouble.

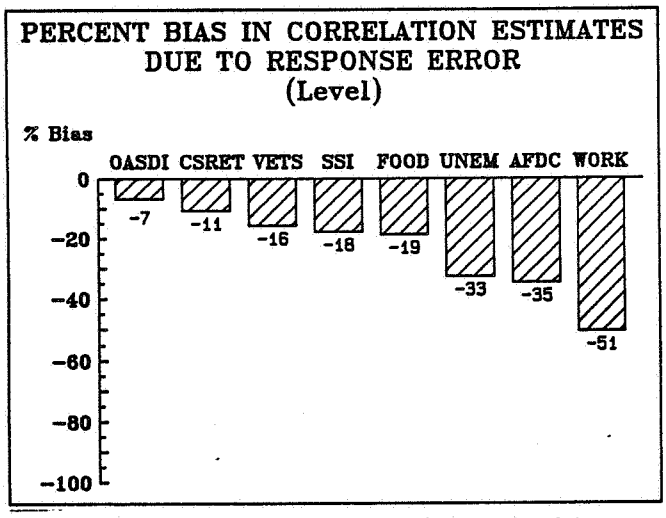

Figure 3.7: For measures of level, biases in estimated correlations due to response errors are trivial for some programs and quite serious for others.

We address the bias in estimated correlations for the change measures in Figure 3.8, looking at the effects on and off the seam separately. Note first that the correlations for all programs are severely biased--the least amount of attenuation is 58 percent for Unemployment Insurance (UNEM) when the measure is off the seam. The biases are more negative for the other programs, reaching -100 percent for the Supplemental Security Income (SSI) estimate on the seam. (We have omitted from the figure two programs that had no true change in at least one pair of months.)

Recall that the on-off seam classification made a big difference in the direction of the biasing effects of error on the estimated mean. Here, however, there is no important effect of the on-off seam classification on the size or sign of the bias in correlation estimates. This is because, while the means of the on and off seam response error distributions have different signs and sizes, the variances of both error distributions are about the same. This is true for each of the programs. And these variances (not shown) are large enough to make it very difficult to detect the true correlational relationships

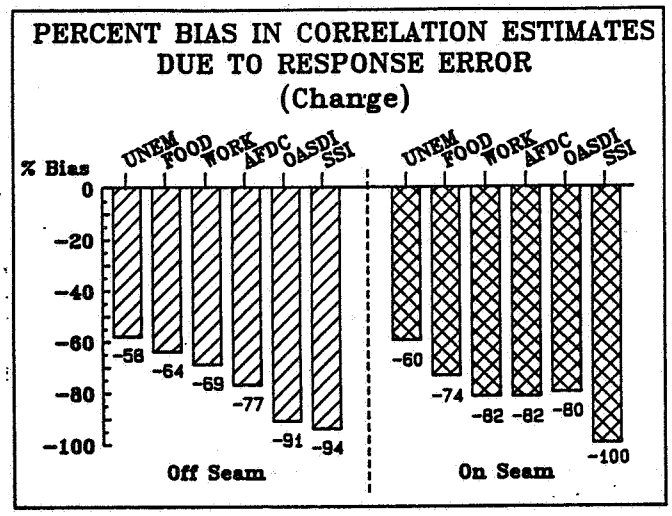

Figure 3.8: For measures of change, correlation biases are consistently very large, regardless of whether they are measured on or off the seam.

in the sample. This also explains why others (e.g., Young, 1989) find that the estimates of correlations and regressions are about the same regardless of whether the change measure is made on or off the seam. The reason is, basically, that correlations and regressions are affected mainly by the second moment or variance of the response error distribution and, in this case, the error variances are about the same relative size.

This concludes our description of the errors in reporting program participation. We have shown that while the errors occur at very low rates, they can have very large effects on the kinds of estimates that analysts want to make from SIPP data. Because response errors have these important effects, we need to understand what is causing them in order to devise strategies for counteracting or removing the causes.

4. CAUSES AND CORRELATES OF THE RESPONSE ERRORS

We will devote the remaining discussion to examining some well-known hypotheses about the causes of response error. These results tend not to support the models of response error implicit in the design of the major government, commercial and academic panel surveys. We cannot say that we have found the correct explanatory model either, but at the end of the discussion we will mention some progress we have made with new research.

We cover five kinds of approaches: the forgetting model, the confusion model, the learning model, the competence model and, at the end, the results of some exploratory cognitive research. As the famous survey methodologist, James N. Morgan, used to say, "There is nothing like real data to give theory a cold bath."

Most hypotheses are formulated in terms of the directional response errors, underreports and/or overreports. Since these errors are conditional on a particular true value, the numbers of cases will vary for each program. We mention the n's in the text.

4.1 The Forgetting Model

The most widely accepted principles for designing factual survey measurements come from the forgetting model. In this model, the response errors are just omission errors (underreports)--people forget that an event occurred. There are few, if any, errors of commission or fabrication (overreports).

Part of the forgetting model is the principle of memory decay over time. As we all know intuitively, the older the to-be-recalled event, the more likely we are to have forgotten it. Therefore, the probability of underreporting should increase as the elapsed time from the event's occurrence increases.

Derivations from the forgetting model have shaped the design of many surveys over the years. We mention three here. First, to minimize forgetting, surveys sometimes use a variety of memory retrieval cues such as many specific questions, each covering a narrow aspect of the topic, and they sometimes use long checklists covering all elements of a topic.

Second, the forgetting model suggests that we use the very shortest recall periods possible to minimize the effects of memory decay forgetting. Although SIPP currently uses a four month recall period, other government surveys use much shorter intervals (e.g., two week recall of illness conditions in the National Health Survey and a one week recall for employment status in the Current Population Survey). Since short reference periods can decrease the precision of survey estimates, the precision must be regained by increasing the sample size (e.g., Gray, 1955), usually a very expensive alternative.

Third, the forgetting model suggests some widely used short cuts for evaluating survey measurements. The best known is the principle of "more is better." Because the

predicted errors are only omission errors, a new survey procedure that gets more reporting of something is a better procedure. Indeed, this theorem means that we never need to do record check studies to evaluate which procedure is better! If we decide to do a record check for some other reason, however, we only need to pay for a one-directional design to measure the underreports (since we can assume that there are few, if any, overreporting errors in the data).

Let us look at how well the forgetting model explains the SIPP response errors in program participation. We will look at the underreporting and overreporting error frequencies for the eight programs and we will look for evidence of memory decay. With one partial exception, we will discover that the errors do not follow the patterns predicted by the forgetting model.

The forgetting model predicts that the response errors will be almost entirely underreporting errors. But looking at the average number of monthly underreports and overreports in Figure 4.1, we see that both the overreporting and underreporting frequencies are substantially greater than zero.4/ And while there are usually more underreports than overreports, the overreport frequencies are often substantially above zero as, for example, is the case for social security (OASDI) where there are 94 overreports in each month on the average and 79 underreports. These kinds of error distributions cannot be explained by a pure forgetting model.

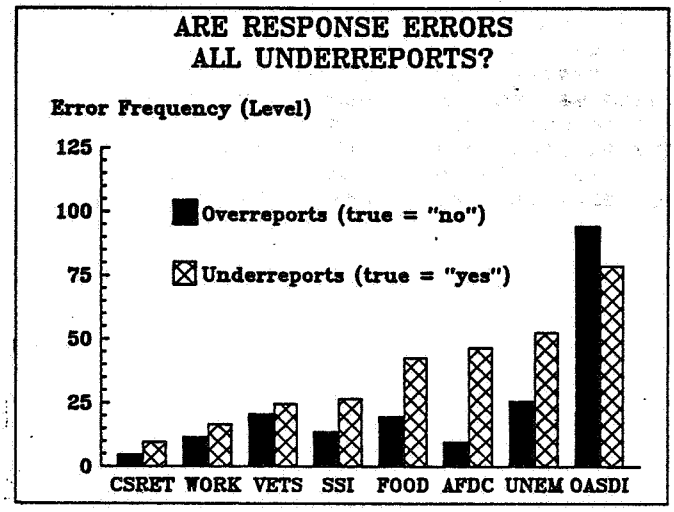

Figure 4.1: Although underreports usually predominate, all programs contain overreports as well.

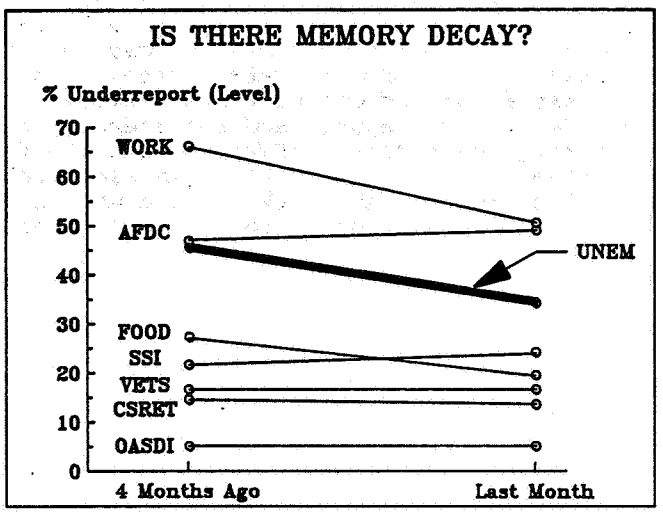

Figure 4.2: Participation underreports for "4 months ago" versus "last month" show little evidence of memory decay.

Before abandoning the forgetting model of causation, however, let us look at its strongest prediction, that error rates follow a time decay pattern. In Figure 4.2 we have plotted the average underreporting rates for participation four months ago and for last month. If memory decay causes the response errors, each line should slope downward. But most of the lines don't slope downward.5/ In only one case, UNEM, is there a meaningful and statistically significant reduction in the underreport rate for the most recent month.

Let us also point out that while the UNEM slope is consistent with the memory decay prediction, the level of error is not. The average underreporting percent in the most recent month for the UNEM program is among the highest observed for any of the eight

---

4/ Based on the standard error of the frequency estimated as $[Np(1-p)]^{1/2}$ where p is the average monthly error probability and N is approximately 7550 for the federally administered programs and 5200 for the state-administered programs.

5/ For each program, the analysis is based on all people who could have underreported (true participation = "yes") either "4 months ago" or "last month" in a wave. Significance testing is for each wave separately, taking account of the within-person correlation of observations over time where appropriate. We report the average underreport percent over waves in Figure 4.2. The t-value for the wave 2 UNEM difference is the only one exceeding 2.00. Numbers of people included in these analyses, by program and wave are: AFDC=111,108 CSRET=69,69 FOOD=215,205 OASDI=1467,1499 SSI=118,121 UNEM=193,203 VETS=149,150 and WORK=42,34.

programs. There is nothing in the pure forgetting models that would predict this; in fact, such models generally assume that recent events are recalled with little or no error at all. The results in Figure 4.2 are contrary to this assumption, and for all programs.

The failure of the memory decay prediction for most programs is the most counter-intuitive result of this research. It is, however, consistent with a growing body of research in autobiographical memory. This finding, in combination with the other research results, has potentially broad implications for SIPP and for survey measurement design in general.

In sum, the response error distributions really don't conform to the fundamental predictions from the forgetting model. The observed underreports do not follow a memory decay forgetting pattern, overreporting is high for many of the programs, and recall of the most recent events is far from error-free. Since there are many errors in both directions, we look next at confusion models. These models predict the generation of both kinds of errors.

4.2 Confusion Models

Confusion models postulate that the underlying trait of interest is reported but that a crucial attribute of the trait is misreported. In record check studies, if someone misreports a feature of an event we may fail to match that event with its true counterpart in the administrative record and, as a result, we may observe one underreport error and one overreport error. A data set that has roughly equivalent frequencies of overreports and underreports could be generated by some process representing respondent confusion about key features used to match the survey and record items. We will look at three dimensions of confusion here: confusion about the name of the program, about the person who is the "official" recipient of benefits, and confusion about the time periods of participation.

4.2.1 Program name confusion.

When we described our preliminary results at last year's conference (Marquis and Moore, 1989a), we mentioned finding confusion about the name for the AFDC program among Pennsylvania respondents; we discovered that many people were reporting their AFDC benefits as General Welfare benefits. Since we didn't check General Welfare records, the confusion hypothesis couldn't be confirmed with certainty, but the finding led us to search for other instances of program name confusion in the remaining programs and states.

The first place we looked was for confusion between the social security (OASDI) and the Supplemental Security Income (SSI) programs. These programs sound alike and either respondents or interviewers may mix them up during the interviews. Indeed, some early research (Vaughan, 1978) indicated that such confusion might be responsible for response errors, and led to a design change to help respondents distinguish among these programs.

Our program name confusion analysis consists of several parts. For a given month for each pair of programs (Program A and Program B) we ask, 1) if the record says the person participated in Program A and didn't participate in Program B, how many times are people reported as not participating in A and participating in B? and 2) the converse, if the record says A=no and B=yes, how often do we observe reports of A=yes and B=no? The significance test (Fisher's exact test) addresses whether there are more such errors than expected. It gives the probability of observing a table with at least as much association as observed when the null hypothesis is true. We illustrate question 1 in Figure 4.3.

We constructed such tables for each of the extreme months (last month and 4 months ago) of wave 1 and wave 2 for each pair of federal programs (and each pair of state programs, but we omit the state data here). The results, averaged over the months, are shown in Table 4.1. When we say that an average frequency is statistically significant, we mean that all of the Fisher exact test p-values (two-tailed) for the four months were less than .10. When we indicate that an average frequency was not significant, we mean that none of the p's were less than .10. (It just worked out that

the test results were consistent among the monthly p values.) The n's for each analysis varied depending on the program, month, and record values.6/

CONFUSION BETWEEN OASDI AND SSI FOR WAVE 1 MONTH 1
CONDITIONAL ON OASDI=YES AND SSI=NO IN THE ADMINISTRATIVE RECORDS

		REPORTED SSI		
		NO	YES	Total
REPORTED OASDI	YES	Both Correct	SSI overreport	OASDI correct
	NO	OASDI Underreport	Name Confusion	OASDI Underreport
	Total	SSI Correct	SSI Overreport	Record OASDI=YES and Record SSI = NO

Figure 4.3: Illustration of program name confusion analysis.

Table 4.1 shows that in the average month, three people who underreported OASDI also made SSI overreports, a rate significantly greater than expected by chance. Similarly, three people who underreported SSI also overreported OASDI. (It is mere coincidence that there were three errors in each direction.) We found an occasional set of "mirrored" errors in other pairs of the federal programs but not at rates that exceeded chance expectations.

TABLE 4.1 AVERAGE MONTHLY FREQUENCIES OF PROGRAM NAME CONFUSION

UNDERREPORT IN:	MATCHED TO OVERREPORT IN: CSRET	OASDI	SSI	VETS
CSRET	-			
OASDI		-	3*	1
SSI		3*	-	
VETS	1	1		-

* p < .05, Fisher's exact test.

To put the results in perspective, consider that, in an average month, there are 79 underreports and 94 overreports of OASDI and there are 27 underreports and 13 overreports of SSI. So name confusion accounts for a small (or null) portion of most errors but potentially accounts for 3/13 = 23 percent of the SSI overreports.

The six cases of mirrored errors in SSI and OASDI are possible instances of program name confusion, although we would be more confident if the other attributes of the misnamed programs were reported correctly, specifically the timing and amounts of the benefits. We examine this in Table 4.2 and conclude that we have gained additional confidence in the conclusion about the SSI underreports being the results of confusion about true OASDI benefit receipt.

In the upper part of Table 4.2 we list the amounts of OASDI benefits underreported in wave 1 and wave 2 on the left, and on the right the amounts of SSI benefits overreported for each case. For the third case, the timing and amounts are close

6/ Numbers of cases used in the program name confusion analyses:

		PROGRAM B (TRUE = NO) OASDI	SSI	VETS	CSRET
PROGRAM A (TRUE = YES)	OASDI	-	1382-1435	1365-1415	1399-1450
	SSI	58-61	-	110-113	113-116
	VETS	76-79	144-147	-	139-142
	CSRET	31-32	69-69	61-61	-

TABLE 4.2: AVERAGE MONTHLY AMOUNT REPORTS FOR POTENTIAL PROGRAM CONFUSION CASES

CASE NUMBER	UNDERREPORTED OASDI AMOUNTS WAVE 1	UNDERREPORTED OASDI AMOUNTS WAVE 2	OVERREPORTED SSI AMOUNTS WAVE 1	OVERREPORTED SSI AMOUNTS WAVE 2	DO AMOUNTS CONFIRM PROGRAM NAME CONFUSION?
1	$182.	$187.	$200.	$247.	Maybe wave 1
2	196.	201.	365.	300.	No
3	391.	402.	387.	387.	Probably
	UNDERREPORTED SSI AMOUNTS		OVERREPORTED OASDI AMOUNTS		
	WAVE 1	WAVE 2	WAVE 1	WAVE 2	
4	379.	413.	872.	414.	Yes, in wave 2
5	304.	312.	304.	314.	Yes
6	0.	314.	0.	314.	Yes

enough to provide a weak subjective confirmation of name confusion. The evidence in the first two cases, however, doesn't add much additional support for the hypothesis. Lack of additional support does not negate the hypothesis since these could be reports that not only confuse the program name but also the benefit amounts and/or timing.

On the other hand, the data in the bottom part of Table 4.2 give additional support to the notion that three SSI underreports are due to confusing the name of the program. The agreement is perfect for the 6th case, close to perfect for the 5th and almost perfect for wave 2 of the 4th case. (Who knows what happened in wave 1? Perhaps the data entry clerk misread the "3" as an "8" in the hundreds position.)

We have extended the name confusion analysis to the state programs also and conclude that there is no additional evidence of program name confusion among the programs in the record check study. (Results not shown; using $p < .10$, we encountered no instance where the number of mirrored errors across pairs of programs was greater than expected.)

Next, we extended the name confusion analysis to include more programs. We are able to do a limited exploration of the hypothesis that names of record checked programs are confused with names of non-record-checked programs. This exploration also used two steps. First, we correlated the underreport scores for each record checked program with rates of reporting non-record-checked programs (there are about 20 such "outside" programs in the analysis). We used these data to answer questions such as whether social security underreporting is related to reports of receiving company pension benefits or state government pension benefits. A positive correlation suggests the possibility of a program name confusion. The second step, for relevant cases in relevant programs, is to examine the reports of benefit amounts and timing, making subjective judgments about whether the additional information supports the name confusion hypothesis.

The results of the analysis (not shown) indicated the potential for confusion only between the checked AFDC program and the unchecked General Welfare programs. (We had detected this result earlier for the state of Pennsylvania and, as we subsequently discovered, so had others (Klein and Vaughan, 1980, and Goudreau, Oberheu, and Vaughan, 1984)). In the second analysis step we found that all of the 30 apparently confused cases resided in Pennsylvania. Of these, 73 percent supported the name confusion hypothesis in that their underreported AFDC benefits timing and amounts agreed closely with their reported timing and amounts of General Welfare benefits.

So, we conclude that the program name confusion model has a limited but useful role in explaining occasional survey reporting errors. It seems to be a major determinant of AFDC underreporting in Pennsylvania and a minor contributor to underreports and overreports of receiving SSI and OASDI benefits.

4.2.2 Person Name Confusion

A second hypothesis, which also predicts both over and underreport errors, is that there is confusion about who is the "official" recipient of the program benefit.

Confusion can arise, for example, when a parent or guardian receives a child's benefit check but the child is the official beneficiary. Or a respondent knows that the family gets Food Stamps but is unsure of whether mother or grandmother is the official recipient in the records.

We define an indicator of recipient name confusion at the household level of analysis. The indicator value is 1 for a given program in a given month if the household contains both an underreport and an overreport error. Otherwise the value of the person name confusion indicator is zero. For our analysis we compare the observed and expected number of instances of recipient name confusion for each month of each program within the subset of households who might confuse the name of the official recipient (those containing someone with a true value of yes and someone else with a true value of no for that program and month).7/ We use Fisher's exact test for inferences about statistical significance.

(Households are defined geographically at the beginning of wave 1 as all eligible persons living in the same dwelling unit. The designation of a person's household does not change over time even if he or she moves out, acquires a different household head, or for any other reason. We have developed a different method for defining the longitudinal family which recognizes such changes and reconfigures households accordingly. However, this procedure is a little cumbersome and costly to execute and results in excluding many cases from those analyses that cannot tolerate partially missing data over time. Our strategy, therefore, is to use the crude household definition for this first look at the hypothesis.)

The results (not shown) indicate that, in the typical month, only the Food Stamps (FOOD) program has significantly more observed than expected instances of person name confusion; typically eight households both underreported and overreported FOOD participation each month. To gain some perspective, note that there are 19 overreports and 43 underreports (62 misclassification errors) for the FOOD program in a typical month. At least 40 percent of the monthly FOOD overreports, 20 percent of the underreports, and at least 25 percent of the FOOD misclassification errors are due to person name confusion.

4.2.3 Time Period Confusion

Our final confusion hypotheses concern the misreporting of the time of program participation. A popular set of hypotheses in the survey methods literature concerns telescoping, the misplacement of the true event in calendar time (e.g., Neter and Waksberg, 1966, Sudman and Bradburn, 1973). Originally, telescoping referred to recalling an event more recently than it truly happened. Subsequently this was labeled forward telescoping to distinguish it from remembering an event further back in time than it truly happened. More recent thinking has introduced the concepts of internal and external telescoping referring to whether the time confusion is merely within the survey reference period (internal) or involves incorrectly placing an event into or out of the reference period (external). Our analyses will include tests of hypotheses about both internal and external telescoping. We find very little support for the telescoping confusion models.

When there is a true change in participation status at the individual level, a respondent who internally telescopes will underreport participation in one month and overreport it in a subsequent month. The implication of this pattern for the whole sample is an increasing overreporting rate (and a decreasing underreporting rate) as one moves from the more distant to the more recent months of the reference period. Since we have already seen (in the memory decay analysis, Section 4.1) that underreport rates seldom show a time effect, we look here at the overreporting percentages, again using the line chart approach. In Figure 4.4 we compare, for each program, the average overreporting percent for 4 months ago with the percent for last month. The averages

---

7/ The expected number is the cell frequency calculated from the marginals under the hypothesis of independence. The ranges for the number of eligible households by program (over months) are: AFDC=46-50, CSRET=37, FOOD=79-86, OASDI=351-368, SSI=43-48, UNEM=101-113, VETS=104-106, WORK=21-29.

are over the two waves.8/ If the forward internal telescoping model fits the data, we should see upward sloping lines.

According to the results in Figure 4.4 for no program do the overreport error differences indicate a statistically significant internal telescoping effect. This result may be due, in part, to low rates of true program participation change for some programs and we cannot rule out the possibility that some individuals may have made internal telescoping errors. However, judging from the temporal pattern of overreport errors, it is clear that internal telescoping is not a major cause of reporting error problems in SIPP program participation.

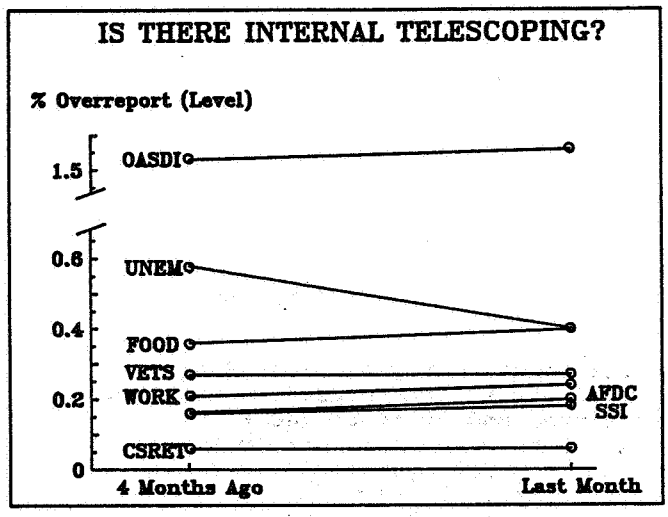

Figure 4.4: Overreports for "4 months ago" vs. "last month" do not show a forward internal telescoping pattern.

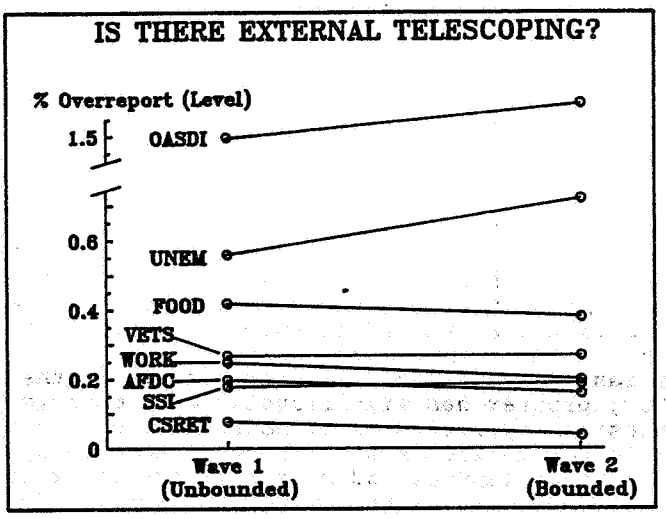

Figure 4.5: External telescoping into an unbounded (wave 1) reference period does not explain observed overreporting either.

Finally, to test the external telescoping hypothesis, we look at the data in Figure 4.5 to see whether the overreporting percent is greater in wave 1 than in wave 2 for each program. The reasoning is that respondents may telescope instances of past program participation into the wave 1 reference period because the start of wave 1 is "unbounded" by a salient event (such as being interviewed by a Census Bureau Field Representative). Telescoping a past participation into wave 2 is unlikely, however, because the beginning of wave 2 is bounded approximately by the experience of the wave 1 interview. A respondent should be able to remember whether a particular participation event happened before or after the last interview and, in addition, remember whether it was reported in the last interview or not.9/ For Figure 4.5 we averaged the overreporting percents over the four months in each wave and used the

---

8/ For each program and each wave, we compare the overreport percents based on all people who could have overreported (true participation = "no") either "4 months ago" or "last month." Significance testing is for each wave separately, taking account of the within-person correlation of observations over time as appropriate. For no program was the within-wave difference statistically significant for either of the two waves. Numbers of people included in these analyses, by program and wave, are: AFDC=5129,5127 CSRET=7478,7478 FOOD=5053,5066 OASDI=6114,6093 SSI=7437,7440 UNEM=5136,5140 VETS=7400,7397 and WORK=5195,5190.

9/ At the suggestion of the discussant, Seymour Sudman, we should mention that the wave 2 SIPP respondent is reminded of which programs were reported for the sample person in wave 1. However, the respondent is not told in which months (of the four-month reference period in wave 1) the sample person was participating. From this definitional perspective, the bounding of the second interview is not necessarily as helpful as it could be. But let us also mention that this kind of "dependent" interviewing is controversial. While it can help a respondent place an event in time, it can also lead to a correlation of response errors over time that might not have happened otherwise.

paired comparison t-test on the averages to make inferences about statistical significance.10/

None of the wave-to-wave overreporting differences in Figure 4.5 is statistically significant using a paired comparison t-test on the averages. Although external telescoping could explain errors made by a few respondents, external telescoping does not predict the pattern of overreporting observed for the sample as a whole.

To summarize, the confusion model is helpful in explaining isolated instances of response errors, such as confusion in Pennsylvania between the AFDC and General Welfare program names, occasional confusion between social security and Supplemental Security Income, and confusion within households about the official recipient for the Food Stamps benefit. Confusion about the timing of participation does not account importantly for the errors. Confusion about attributes, then, is not a broadly useful principle for understanding and fixing SIPP response errors.

4.3 Learning Models

Another class of explanations postulates that people learn to change their behavior and this affects the quality of data obtained in panel or longitudinal surveys. We examine the hypotheses (1) that people learn to underreport over time and (2) that people deliberately change their true program participation behavior as a result of being interviewed. Neither learning hypothesis receives much support.

While we hope that respondents will learn to report more accurately and fully over repeated panel interviews, the more commonly heard hypothesis is cynical: respondents will learn to underreport the events of interest in subsequent interviews in order to avoid the long, tedious set of questions about details of the reported target events. In SIPP there are tedious, difficult questions about exact amounts and timing of benefits for programs in which the sample person is participating. By not reporting program participation, the respondent can avoid the unpleasant questions.

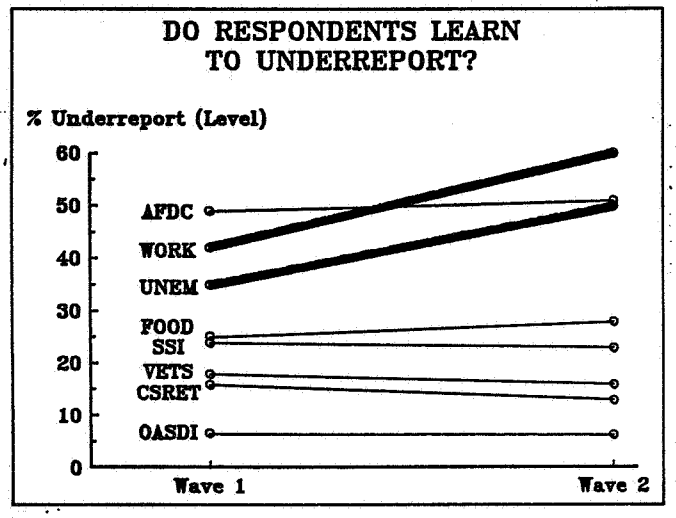

Figure 4.6: By wave 2, respondents may have "learned" to underreport UNEM and WORK participation.

In Figure 4.6 we look at the change in underreporting rates between wave 1 and wave 2. If respondents learn to avoid unpleasant questions, they will underreport more in wave 2 compared to wave 1. Again, we average the monthly underreport percents for each wave and apply the paired comparison t-test to the averages.11/

Two of the eight programs shown in Figure 4.6 have the predicted upward sloping lines, Workers' Compensation (WORK) and Unemployment Insurance (UNEM). For the remaining programs the differences are not statistically significant. While the size of the significant effects in the predicted direction is small, we are inclined to take them seriously. With the benefit of hindsight we note that the reconstruction of the amounts and timing of Unemployment benefits and Workers Compensation benefits can be extremely difficult. Benefits from these programs are usually not paid as regularly as those of the other transfer programs, payments are made for periods of less than a month, and the amounts can vary considerably from month to month. Perhaps we should not be surprised that some

---

10/ Cases included in the analysis are those whose administrative record participation value was "no" at least once in each wave. The n's for the comparison of the averages are: AFDC=5115, CSRET=7478, FOOD=5032, OASDI=6079, SSI=7433, UNEM=5104, VETS=7397, and WORK=5185.

11/ To be included in the analysis for a program, the administrative record needed to indicate that the sample person truly participated at least once in wave 1 and once in wave 2. The N's on which the comparisons are based are: AFDC=98, CSRET=69, FOOD=186, OASDI=1458, SSI=117, UNEM=135, VETS=149, WORK=27.

respondents would rather not attempt this difficult reconstruction in more than one interview.12/

4.3.2 Reactive Measurement Effects

Another learning issue for designers of panel surveys is related to the Heisenberg Principle: the act of measurement distorts the phenomenon being measured. In SIPP we ask whether interviewing people about their participation in welfare programs causes them to subsequently enroll in the programs, something they might not have done if they had not been interviewed. We look at this in Figure 4.7 by comparing average true participation rates in wave 1 with average true rates in wave 2. We use all available cases and base our inferences on paired comparison t-tests.

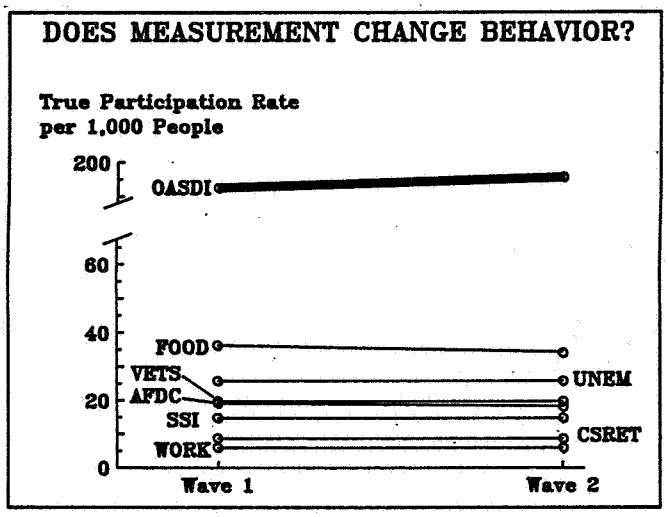

Figure 4.7: There is almost no evidence that respondents increase their program participation as a result of the wave 1 interview.

The data indicate that, in this study, only one program, social security (OASDI), showed a significant increase in average participation from wave 1 to wave 2. However, since social security is a very well known program, the change is probably due to the natural aging of the sample into eligibility rather than to any reactive effect of the wave 1 measurement. (Note that sample persons who died during the survey period--and hence stopped receiving OASDI benefits--are excluded from all analyses in this paper. Thus, sample aging does not produce a corresponding decrease in OASDI participation.)

So, for only two programs do we find evidence that respondent learning creates additional response error or measurement problems over time.

4.4 Competence Models

We label this final group of hypotheses the competence models because they deal with the abilities of the interviewer and respondent to furnish error-free data. We look first at interviewer effects and then at the characteristic of self-proxy respondent status.

4.4.1 Interviewer Effects

To examine interviewer effects we estimate the proportion of the error variance that is contributed by variation among interviewers. Our basic measurement model is:

$$Yij = \beta Xij + Ei + Wij,$$

where:

Yij is the misclassification score (0 if correct response, 1 if wrong),
X represents a vector of sample person characteristics (e.g., age, sex),
β is a vector of regression coefficients,
Ei is the effect of interviewer i (mean = 0, variance = Var E),
Wij is the sample person specific effect (variance = Var W), and
i indexes the interviewer and j indexes the sample person.

We estimate a random effects regression model for the interviewer term and treat the other variables as fixed. We express the interviewer effect, rho, as the ratio of the variances: Var E / (Var E + Var W). Because interviewers were not allocated to respondents using an interpenetrated design, we include the vector of respondent characteristic variables to account for differences among respondents that could be confounded with the interviewer assignments. Also, at this preliminary stage, we have not estimated the interviewer effect for every month of every program. Instead, we made the estimates for two extreme time periods (wave 1, four months ago and wave 2, last month). Figure 4.8 shows the average of the two estimates for each program. The

12/ See Seymour Sudman's discussion for a different interpretation involving the social stigma associated with chronic unemployment.

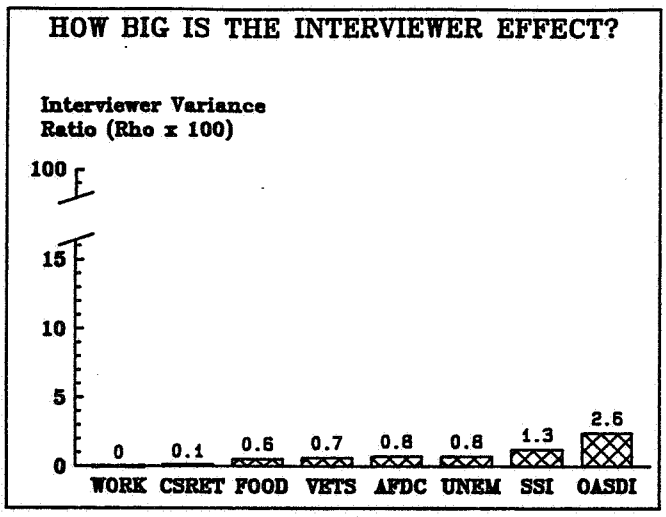

Figure 4.8: The interviewers' contribution to the response error variance is around 1 or 2 percent.

number of interviewers in the monthly estimates varies between 92 and 109.

In the survey methods literature (e.g., Fowler and Mangione, 1990) one usually sees estimates of rho in the one-to-two percent range and this is about what we found here. Figure 4.8 suggests that, averaging rho's over months, approximately one percent of the response error is due to idiosyncratic ways that interviewers collect the data. So, these interviewer effects are not a major source of response error variation in SIPP. And, unless interviewer is confounded with a variable that an analyst uses for classification, these interviewer effects are not going to distort subject matter estimates importantly.

4.4.2 Effects of Self-Proxy Respondent Status

The conventional wisdom asserts that respondents are more competent if they are reporting about themselves than when they are reporting for someone else (being a proxy respondent). Moore's (1988) literature review has called this into question because few of the studies that found more errors in proxy responses had randomized the respondent status treatment in an experimental design. Such studies did not attempt to rule out other causes of the errors that might be confounded with the naturally occurring respondent status. It is important to know if proxy response is really much worse than self response because implementing a self response rule in a national panel survey can be very expensive.

Our record check study also lacks the randomized treatment design of a proper experiment so self/proxy status may be confounded with other variables and self/proxy effects on error rates may reflect only the effects of these other variables.

In Table 4.3 we show self/proxy effects on all three types of response error: misclassifications, underreports, and overreports. Within a household, self/proxy status can change from one interview to another so we estimate self/proxy effects separately for each wave. We average the monthly error rates over the four months covered by the interview and test the significance of the difference between the mean self report error rate and the mean proxy report error rate, using an approximate t-test procedure that adjusts for unequal variances and that uses the Satterthwaite approximation for the degrees of freedom. The parenthetical entries in the (n) rows indicate the numbers of cases in each respondent status class for each type of error. Because these are averages over four months, a person will appear in both the analysis of overreports and underreports if the administrative record indicates both a yes and no participation in the program for the period (so the sum of the underreport and overreport n's may exceed the n in the misclassification analysis).

As shown in Table 4.3, when the dependent variable is the misclassification error (first 3 data columns), the self/proxy effects are statistically significant in both waves for only the OASDI program (we use a double underline of the difference value to indicate statistical significance). Perhaps of some interest is the fact that in all of the eight programs, self responses contained more misclassification errors than proxy responses in at least one wave. This may be because the probability of a wrong answer is higher when the true value is yes and because the true value is more likely to be yes for self respondents (an example of confounding).

So we controlled for true participation by modeling underreporting and overreporting separately. The underreporting results in Table 4.3 show consistently (both waves) and significantly higher underreporting percentages for proxy respondents in only one of the eight programs, AFDC. However, in six of the programs, the consistent direction of the effect is for proxy responses to contain the most underreport errors. The overreport results resemble the misclassification results (mixed signs, only OASDI shows statistically significant differences in both waves, indicating that self-respondents make more overreporting errors than proxy respondents).

We are refining our analysis model at the present time. In the current specification, we attempt to explain both overreporting and underreporting separately and we introduce

TABLE 4.3: PERCENT RESPONSE ERROR BY SELF/PROXY RESPONDENT STATUS, PROGRAM, AND WAVE

		Missclassification			Underreport			Overreport		
PROGRAM	WAVE	Self	Proxy	Diff	Self	Proxy	Diff	Self	Proxy	Diff
AFDC	1	.012	.009	+.003	.47	.77	-.30	.002	.002	+.001
	(ave. n)	(3541)	(1671)		(94)	(22)		(3471)	(1659)	
	2	.013	.007	+.006	.48	.75	-.27	.002	.002	+.000
	(ave. n)	(3400)	(1812)		(91)	(19)		(3325)	(1802)	
FOOD	1	.014	.009	+.005	.25	.36	-.11	.005	.004	+.002
	(n)	(3541)	(1671)		(184)	(36)		(3407)	(1648)	
	2	.013	.009	+.004	.24	.48	-.24	.006	.004	+.002
	(n)	(3400)	(1812)		(173)	(35)		(3277)	(1793)	
UNEM	1	.013	.018	-.006	.41	.54	-.13	.006	.009	-.004
	(n)	(3539)	(1669)		(151)	(68)		(3490)	(1657)	
	2	.016	.016	+.000	.43	.51	-.08	.009	.008	+.001
	(n)	(3397)	(1811)		(153)	(74)		(3358)	(1795)	
WORK	1	.005	.007	-.002	.45	.73	-.28	.002	.004	+.001
	(n)	(3540)	(1672)		(32)	(15)		(3527)	(1668)	
	2	.005	.005	+.000	.57	.75	-.18	.002	.001	+.001
	(n)	(3399)	(1813)		(28)	(12)		(3384)	(1806)	
CSRET	1	.003	.001	+.002	.16	.17	-.01	.001	.001	+.000
	(n)	(5139)	(2408)		(63)	(6)		(5076)	(2402)	
	2	.002	.001	+.001	.12	.29	-.17	.0004	.0004	+.000
	(n)	(4894)	(2653)		(62)	(7)		(4832)	(2646)	
OASDI	1	.025	.017	+.008	.06	.10	-.04	.018	.009	+.009
	(n)	(5140)	(2410)		(1230)	(238)		(3935)	(2179)	
	2	.027	.017	+.010	.06	.08	-.03	.021	.010	+.011
	(n)	(4895)	(2655)		(1186)	(313)		(3744)	(2349)	
SSI	1	.006	.005	+.001	.26	.24	+.02	.002	.002	-.001
	(n)	(5139)	(2409)		(90)	(29)		(5054)	(2383)	
	2	.006	.004	+.002	.27	.20	+.07	.002	.002	-.001
	(n)	(4894)	(2654)		(91)	(30)		(4812)	(2628)	
VETS	1	.007	.005	+.002	.18	.17	+.01	.003	.002	+.002
	(n)	(5139)	(2408)		(107)	(42)		(5034)	(2366)	
	2	.006	.006	+.000	.16	.18	-.02	.003	.003	+.000
	(n)	(4894)	(2653)		(105)	(45)		(4789)	(2608)	

other predictor variables in order to remove confounding that would ordinarily be controlled in an experimental design.

However, regardless of the outcome of the modeling of self/proxy response error differences, the practical implications will be the same: because self response error levels are so high, instituting a self response rule in SIPP may reduce errors a little, but not nearly enough. It will be more important to teach the respondent how to respond well throughout the entire panel survey, both as a self respondent and as a proxy for someone else.

4.5 Supplementary Studies

Before concluding we will mention two additional studies conducted to yield additional

insights into the record check results: an experimental evaluation of recall decay and some exploratory research on the cognitive processes respondents use in SIPP. We mention only key results here since we hope to present each of these studies in detail elsewhere.

4.5.1 Experimental Study of Recall Decay

In the present study the memory decay model does not appear to explain the observed pattern of reporting errors, clearly contrary to conventional wisdom. But other recent research also finds that memory decay is not a major predictor of short-term omission errors. For example, a few years ago we conducted a small randomized experiment with a 10 percent sample of Census Bureau headquarters employees. Using a self-administered questionnaire, we asked them to recall their use of sick leave and vacation leave. We asked half the sample to recall over a recent three month period and the other half over a recent six month period. In addition we asked everyone to recall vacation and sick leave for the last complete calendar year (which ended nine months before we took the survey). Using administrative leave records to evaluate the responses, we found no effects of the elapsed time for vacation time response errors. There were plenty of response errors, but neither their size, direction, nor frequency was affected by the various recall lengths. For sick leave, however, the response error rates were a little greater for the longest recall interval although the three- vs. six-month treatment had no detectable effect.

Several other studies of autobiographical memory (e.g. Linton, 1986, and Wagenaar, 1986) also suggest little or no memory decay effect during an event's first year in memory and very small decay effects, perhaps only five percent per year, following that. The point is that since research in other settings has cast doubt on the critical role of memory decay, it is also conceivable that memory decay is not an important cause of SIPP response errors.

4.5.2 Exploratory Cognitive Research

Since we have been unsuccessful in uncovering the major causes of SIPP response errors, we began some exploratory cognitive research last summer aimed at a better understanding of respondents' thought processes in answering SIPP questions. We provided basic training to half a dozen headquarters staff members in techniques for eliciting thinking processes during interviews. These staff members accompanied experienced SIPP interviewers to nonsample households in the headquarters area. Staff members interrupted the interviews at appropriate places to learn whatever the respondent could reveal about the cognitive answering processes. The data for this research are the staff members' written summaries of the important verbal interactions which occurred during each tape-recorded interview. Although we did not observe many welfare program recipients, we did observe reporting of similar regular and irregular income streams and feel that the results may generalize.

One of the main conclusions from our review of the written summaries is that many respondents adopt a very simple heuristic or rule of thumb to quickly answer any question about a specific four-month stream of income. They use the simple rule as a substitute for detailed, direct recall and a substitute for checking their personal records. These simple heuristics do not necessarily bias mean estimates, since they seem just as likely to yield overreports as underreports, but they tend to obscure the real details, changes and other phenomena that transpired over the time period, adding a considerable amount of "noise" to the observations. As a result of the exploratory research, we are formulating hypotheses about how to preempt the initial use of simple heuristics and how to teach respondents the correct problem solving strategies for accurate financial reporting.

5. CONCLUSIONS

In this section we review our results and discuss what they may imply.

We began by showing that response errors are very rare in SIPP: generally, less than two percent of the answers about program participation or program participation change are wrong. But, because true participation and true change are also rare, the response errors cause important biases in estimates made by SIPP data users. Levels of participation can be underestimated by 10 to 40 percent for many programs; change rates can be underestimated by even greater amounts if the changes are measured off the interview seam, and change rates can be severely overestimated using change data from the interview seams.

It might be argued, however, that these are not important errors since SIPP data are not often used to make estimates of level. Rather, SIPP data are most likely to be used to estimate relationships among variables such as in tests of hypotheses about the causes of going on or off welfare or fitting multivariate policy models of household economic behavior.

Unfortunately, the record check study shows that the SIPP response errors have even larger biasing effects on such relationship estimates. For correlations involving SIPP change measures, biases range from -50 percent to -100 percent. For correlations involving SIPP measures of level, the biases are in the -10 to -50 percent range.

Because these levels of response errors are important, SIPP will want to do something about them. We discuss the two basic options next, informing users about existing errors and minimizing future errors.

In the near future, it will be desirable to inform users as fully as possible about response errors so analysts can try to account for them when making estimates. This would supplement information about response error in existing SIPP error profiles (e.g., King, Petroni, and Singh, 1987; Committee on National Statistics, 1989; Jabine, King, and Petroni, 1990). The best way to achieve this goal is for SIPP to start a continuing administrative record program that would produce estimates of response errors and error covariances on a historical basis. The design of the program could advance well beyond what the present study has done, providing more representative household samples, more sources of administrative records, and allowing estimates of the sizes of other error sources such as attritions, refusals, not-at-homes, movers, etc. The additional information, in a summary form that maintains confidentiality, would be useful to analysts who have the resources to use it.

In the long run, however, the most desirable approach is to reduce errors to much lower levels. Since the 1984 panel (which this record check study evaluated), SIPP has done a number of things to reduce response errors, such as changing the wording of questions, adding items to resolve potential confusion, providing special interviewer training, and testing a time-line calendar procedure. Without record checks, however, it has been difficult to evaluate the effects of these changes.

The current record check results can shed some light on evaluation issues and perhaps move our thinking forward. For example, at one point, some felt that if a procedure reduced the difference in amounts of change reported on and off the seam (see Figure 3.5), the procedure would have reduced the important response errors. However, as we show (Figure 3.6), there can be very large estimation biases in both on and off seam measures. Making one set more like the other is unlikely to help and may even make things worse for measures of level (see Figure 3.3). Others have felt that if reporting of past events could be made as good as the reporting of recent events, then the major response biases would be reduced. Unfortunately, the record check shows that for most programs underreporting and overreporting are about as bad for recent participation as for past program participation (see Figures 4.2 and 4.4). And for the same reason, a procedure that uses reports of recent events as an "anchor" for reports of more remote events is unlikely to result in better data either.

The argument that we developed in last year's Annual Research Conference paper (Marquis and Moore, 1989a) was that to reduce the biases in estimates of level as well as change, one must reduce the response errors in each of the months of the reference period, the most recent as well as the most remote. Indeed, the analyses in this year's paper are simple tests of some of the classical ideas about the causes of response errors, under the assumption that, if we know the causes, we are more likely to find effective solutions quickly. Unfortunately, the tests of the classical hypotheses did not reveal any basic causes underlying all or most--or even a large part--of the errors. They did reveal a few isolated phenomena, such as confusion about the name of the AFDC program in Pennsylvania, confusion between OASDI (social security) and Supplemental Security Income, confusion about the name of the official recipient of Food Stamps, and the possibility that respondents learn to underreport Unemployment Insurance and Workers' Compensation benefits to avoid detailed questioning. Each of these problems has a potential procedural solution that skilled field managers can develop, test and implement.

Looking over our findings we can say that we do not yet know the causes of SIPP response errors or how to fix them. However, our research results, based on a nonexperimental design, suggest to us what might happen if we were to conduct error-reducing experiments based on traditional assumptions about the causes of response errors. They suggest to us that experimenting with providing additional memory cues, rewording questions, reducing interviewer variance, shortening the recall period,

adopting a strict self response rule, reminding the respondent of previously reported information, and similar traditional approaches may not bring the response errors and estimation biases down to satisfactory levels.

Through some exploratory cognitive research we have discovered that SIPP respondents often use very inappropriate cognitive strategies for reconstructing historical information about income streams and program participation. We now feel that these inappropriate problem solving strategies underlie many of the observed response errors in SIPP. While we have a general understanding of why people use such heuristic strategies, the literature is unclear about what to do about it in the household interview setting. (For a different view, see the paper by Krosnick in these proceedings.)

This leads us to our second major recommendation: that SIPP undertake a new program of research to learn how to preempt the use of inappropriate cognitive strategies and encourage respondents to learn and use better ones. A key element will undoubtedly include learning how to provide interviewers with the tools and training necessary to motivate accurate reporting. We view this as a long term research program aimed at uncovering principles and developing procedures applicable to many surveys that impose heavy recall and information processing demands on respondents. We suggest that a wide range of behavioral scientists be invited to contribute their expertise including not only cognitive theorists but also applied specialists in training, motivation, and persuasion.

We recommend that SIPP establish a "field laboratory" in a state that will provide convenient access to income and program participation records for validation purposes. The laboratory would enable developing, testing and evaluating new SIPP interview procedures in households not part of the national sample. The field laboratory facilities would include portable computers so that new versions of the questionnaire can be tested and processed quickly. It should be staffed with skilled interviewers who can adapt easily to new questionnaires and their administration in households with the aid of the computer. And we recommend that SIPP put sufficient priorities and resources into the research to assure implementing major research results in the field by the middle of the next decade (1995-1996). Thereafter, we recommend that SIPP adopt procedures to continuously monitor data quality and to make necessary changes on a continuing basis to keep the response and nonresponse errors within realistic control limits. The control limits should be narrow enough to allow the main analytic uses of the survey data to proceed relatively trouble-free.

REFERENCES

BURKHEAD, D. and CODER, J. (1985), "Gross Changes in Income Recipiency from the Survey of Income and Program Participation," Proceedings of the Social Statistics Section, American Statistical Association, pp. 351-356.

COMMITTEE ON NATIONAL STATISTICS (1989), The Survey of Income and Program Participation: An Interim Assessment, National Academy Press, Washington DC.

FELLEGI, I. and SUNTER, A. (1969), "A Theory for Record Linkage," Journal of the American Statistical Association, Vol. 64, pp. 1183-1210.

FOWLER, F. and MANGIONE, T. (1990), Standardized Survey Interviewing: Minimizing Interviewer-Related Error, Sage Publications, Newbury Park, CA.

GOUDREAU, K., OBERHEU, H. and VAUGHAN, D. (1984), "An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program," Journal of Business and Economic Statistics, Vol. 2, pp. 179-186.

GRAY, P. (1955), "The Memory Factor in Social Surveys," Journal of the American Statistical Association, Vol. 50, pp. 344-363.

HILL, D. (1987), "Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods." Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 210-215.

JABINE, T., KING, K. and PETRONI, R. (1990), "Survey of Income and Program Participation: Quality Profile," U. S. Census Bureau, Washington DC, Forthcoming June 1990.

JARO, M. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, Vol. 84, pp. 414-420.

KING, K., PETRONI, R. and SINGH, R. (1987), "Quality Profile for the Survey of Income and Program Participation," SIPP Working Paper No. 8708, U. S. Census Bureau, Washington DC.

KLEIN, B. and VAUGHAN, D. (1980), "Validity of AFDC Reporting Among List Frame Recipients," Chapter 11 in Olson, J. (ed.), Reports from the Site Research Test, U.S. Department of Health and Human Services, ASPE/ISDP/SIPP, Washington, DC.

LaPLANT, W. (1989), "Users' Manual for the Generalized Record Linkage Program Generator," Statistical Research Division, U.S. Census Bureau, Washington, DC.

LINTON, M. (1986), "Ways of Searching and the Contents of Memory," In D. C. Rubin, Autobiographical Memory, Cambridge University Press, Cambridge.

MARQUIS, K. (1978), "Inferring Health Interview Response Bias from Imperfect Record Checks," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 265-270.

MARQUIS, K. and MOORE, J. (1989a), "Response Errors in SIPP: Preliminary Results," Proceedings of the Fifth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 515-536.

MARQUIS, K. and MOORE, J. (1989b), "Some Response Errors in SIPP--with Thoughts About Their Effects and Remedies," Proceedings of the Section on Survey Research Methods, American Statistical Association, (forthcoming).

MOORE, J. (1988), "Self/Proxy Response Status and Survey Response Quality--A Review of the Literature," Journal of Official Statistics, Vol. 4, pp. 155-172.

MOORE, J. and KASPRZYK, D. (1984), "Month-to-Month Recipiency Turnover in the ISDP," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 210-215.

MOORE, J. and MARQUIS, K. (1989), "Using Administrative Record Data to Evaluate the Quality of Survey Estimates." Survey Methodology, Vol. 15, 129-143.

NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985), "An Overview of the Survey of Income and Program Participation, Update 1." SIPP Working Paper Series, No. 8401, Washington, DC: U.S. Bureau of the Census.

NETER, J. and WAKSBERG, J. (1966), " A Study of Response Errors in Expenditures Data From Household Interviews," Journal of the American Statistical Association, Vol. 59, pp. 18-55.

SUDMAN, S. and BRADBURN, N. (1973), "Effect of Time and Memory Factors on Response in Surveys," Journal of the American Statistical Association, Vol. 68, pp. 805-815.

VAUGHAN, D. (1978), "Errors in Reporting Supplemental Security Income Recipiency in A Pilot Household Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 288-293.

WAGENAAR, W. (1986), "My Memory: A Study of Autobiographical Memory Over Six Years," Cognitive Psychology, Vol. 18, pp. 225-252.

YOUNG, N. (1989), "Wave Seam Effects in SIPP," Proceedings of the Section on Survey Research Methods, American Statistical Association, (Forthcoming).

APPENDIX

Here, we derive the effects of response error on the correlation estimate using a classical measurement model and offer comments about an expanded model. Let us begin with the classical model:

$$\text{Let } M = T + e,$$

where M is the measured response, T is the true value and e is the response error. For 0,1 variables, e is a linear function of true values and a random variable, u, such that:

$$e = \alpha + \beta\, T + u.$$

The expected value of u is zero. β is a parameter representing the degree that errors are correlated with true values. For dichotomous variables, this correlation is negative when any response error is present.

Define Z as a perfectly measured variable. Without loss of generality, define its mean as zero and its scores as deviations from the mean. Also $\mathrm{CovM,Z} = (1 + \beta)\ \mathrm{CovT,Z}$.

The Pearson product-moment correlation, r, between true participation, T, and a perfectly measured variable whose values are deviation scores, Z, is

$$r = \mathrm{CovT,Z} / (\mathrm{VarT\ VarZ})^{.5}$$

and the correlation, r', using measured participation, M, is

$$\begin{aligned} r' &= \mathrm{CovM,Z} / (\mathrm{VarM\ VarZ})^{.5} \\ &= [(1 + \beta)\ \mathrm{CovT,Z} / (\mathrm{VarM\ VarZ})^{.5}]\ (\mathrm{VarT\ VarT})^{.5} \\ &= (1 + \beta)\ (\mathrm{VarT\ VarM})^{.5}\, r. \end{aligned}$$

The bias in the correlation estimate using measured values relative to the correlation using true values, RB(r'), is:

$$\begin{aligned} \mathrm{RB}(r') &= (r - r') / r \\ &= [(1 + \beta)\ (\mathrm{VarT} / \mathrm{VarM})^{.5}\, r - r] / r \\ &= (1 + \beta)\ (\mathrm{VarT} / \mathrm{VarM})^{.5} - 1. \end{aligned}$$

We multiply RB(r') by 100 to express it as a percent.

A reviewer pointed out that in this model the response errors depend on the true value of the current time period and that attenuation in the correlation estimate could be alleviated if response errors also depend on true values in other time periods and the value of the Z variable does not change over the extended time period. As a result we investigated a model in which:

$$e = \alpha + \beta\, T + \beta'\, T' + u,$$

where β' is a parameter representing the degree of correlation between current period response error and the true value in the previous month (T'). We estimated β' for a sample of 3 time periods for all programs. Typically β' was small. About half of the estimates were statistically significant and all of the non-zero results were positive. For federally administered programs, the estimates of correlation attenuation from the alternative model were between 90 and 100 percent of the estimates from the original model. For state-administered programs, the estimates of correlation attenuation from the alternative model were between 80 and 100 percent of the estimates from the original. Since the alternative model suggests slightly smaller amounts of bias in the correlation estimates, models with additional terms may show even less attenuation.

APPENDIX TABLE 1: Crosstabulation of SIPP and Administrative Record Program Participation Reports by Program, Wave, and Month

				SIPP Reference Month							
				4 mos. ago		3 mos. ago		2 mos. ago		last month	
		Record:		no	yes	no	yes	no	yes	no	yes
State Programs:											
AFDC	Wave 1	SIPP:	no	5111	43	5103	50	5104	49	5101	50
			yes	8	50	9	50	10	49	10	51
	Wave 2	SIPP:	no	5105	48	5109	45	5109	45	5107	46
			yes	8	51	9	49	9	49	11	48
FOOD	Wave 1	SIPP:	no	5003	52	5003	48	5008	43	5008	41
			yes	21	136	18	143	19	142	18	145
	Wave 2	SIPP:	no	5012	49	5012	45	5022	35	5024	31
			yes	15	136	22	133	18	137	22	135
UNEM	Wave 1	SIPP:	no	5034	64	5044	54	5056	45	5067	42
			yes	30	80	28	82	18	89	20	79
	Wave 2	SIPP:	no	5042	64	5038	55	5052	53	5053	46
			yes	29	73	37	78	19	84	21	88
WORK	Wave 1	SIPP:	no	5171	18	5168	16	5170	13	5167	14
			yes	13	10	13	15	11	18	14	17
	Wave 2	SIPP:	no	5172	21	5169	20	5170	17	5177	14
			yes	9	10	10	13	9	16	10	11
Federal Programs:											
CSRET	Wave 1	SIPP:	no	7472	11	7472	11	7472	11	7473	11
			yes	6	58	6	58	6	58	5	58
	Wave 2	SIPP:	no	7475	10	7475	10	7475	9	7475	8
			yes	3	59	3	59	3	60	3	61
OASDI	Wave 1	SIPP:	no	6021	76	6012	83	6001	80	5993	78
			yes	90	1363	89	1366	91	1378	93	1386
	Wave 2	SIPP:	no	5988	77	5981	77	5973	78	5962	83
			yes	97	1388	97	1395	95	1404	97	1408
SSI	Wave 1	SIPP:	no	7421	25	7419	27	7418	28	7419	27
			yes	14	88	14	88	13	89	13	89
	Wave 2	SIPP:	no	7423	25	7421	26	7420	27	7420	28
			yes	10	90	13	88	13	88	14	86
VETS	Wave 1	SIPP:	no	7380	25	7379	25	7379	25	7378	26
			yes	20	122	20	123	20	123	20	123
	Wave 2	SIPP:	no	7377	25	7377	25	7377	25	7377	25
			yes	20	125	20	125	20	125	20	125

APPENDIX TABLE 2: Crosstabulation of SIPP and Administrative Record Program Participation Change by Program, Wave, and Month-Pair

				SIPP Reference Month-Pairs							
				mo.4 - mo.3		mo.3 - mo.2		mo.2 - mo.1		"SEAM"	
		Record:		no	yes	no	yes	no	yes	no	yes
State Programs:											
AFDC	Wave 1	SIPP:	no	5197	14	5196	12	5201	9	5192	8
			yes	0	1	2	2	0	2	10	2
	Wave 2	SIPP:	no	5198	9	5203	5	5197	10		
			yes	5	0	3	1	3	2		
FOOD	Wave 1	SIPP:	no	5178	16	5179	13	5182	16	5162	10
			yes	13	5	7	13	5	9	33	7
	Wave 2	SIPP:	no	5178	20	5176	24	5178	14		
			yes	11	3	6	6	12	8		
UNEM	Wave 1	SIPP:	no	5132	36	5109	50	5123	37	5073	40
			yes	20	20	19	30	20	28	61	34
	Wave 2	SIPP:	no	5110	53	5106	52	5116	42		
			yes	26	19	22	28	29	21		
WORK	Wave 1	SIPP:	no	5192	7	5195	6	5187	9	5186	8
			yes	9	4	7	4	11	5	16	2
	Wave 2	SIPP:	no	5200	6	5200	6	5197	5		
			yes	4	2	6	0	5	5		
Federal Programs:											
CSRET	Wave 1	SIPP:	no	7547	0	7547	0	7546	0	7542	0
			yes	0	0	0	0	1	0	5	0
	Wave 2	SIPP:	no	7547	0	7546	0	7546	0		
			yes	0	0	1	0	1	0		
OASDI	Wave 1	SIPP:	no	7527	11	7522	12	7525	11	7493	13
			yes	11	1	15	1	13	1	38	6
	Wave 2	SIPP:	no	7527	12	7525	14	7522	16		
			yes	10	1	9	2	11	1		
SSI	Wave 1	SIPP:	no	7544	4	7542	4	7545	1	7537	1
			yes	0	0	2	0	2	0	10	0
	Wave 2	SIPP:	no	7540	5	7544	2	7542	5		
			yes	3	0	1	1	1	0		
VETS	Wave 1	SIPP:	no	7545	1	7547	0	7546	1	7540	1
			yes	1	0	0	0	0	0	6	0
	Wave 2	SIPP:	no	7547	0	7547	0	7547	0		
			yes	0	0	0	0	0	0		